IIIT_DWD@EACL2021: Identifying Troll Meme in Tamil using a hybrid deep learning approach

Ankit Kumar Mishra[1] and **Sunil Saumya**[2]
[1]Magadh University, Bodh Gaya, Bihar, India
[2]Indian Institute of Information Technology Dharwad, Karnataka, India
ankitmishra.in.com@gmail.com, sunil.saumya@iiitdwd.ac.in

Abstract

Social media are an open forum that allows people to share their knowledge, abilities, talents, ideas, or expressions. Simultaneously, it also allows people to post disrespectful, trolling, defamation, or negative content targeting users or the community based on their gender, race, religious beliefs, etc. Such posts are available in the form of text, image, video, and meme. Among them, memes are currently widely used to disseminate offensive material amongst people. It is primarily in the form of pictures and text. In the present paper, *troll memes* are identified, which is necessary to create a healthy society. To do so, a hybrid deep learning model combining convolutional neural networks and bidirectional long short term memory is proposed to identify trolled memes. The dataset used in the study is a part of the competition *EACL 2021: Troll Meme classification in Tamil*. The proposed model obtained 10th rank in the competition and reported a precision of 0.52, recall 0.59, and weighted F1 0.3.

1 Introduction

Social media has become one of the essential components of our life, thanks to digitization. Over the years, the number of social media users has grown at an exponential pace. According to the survey of *wearesocial*[1], out of around 7.8 billion people worldwide, 3.8 billion people are active social media users (Chakravarthi et al., 2020c; Mandl et al., 2020). They participate in many ways on social media, like posting comments (in the form of text, image, video, audio), online chatting (text, audio, and video), hosting live events, and so on (Chakravarthi et al., 2020b,a). The information available in these posts is overwhelming and very difficult to manage for users. Therefore, people always look for summarized information available on social media instead of reading the information in the form of text paragraphs. Internet memes are one such summarized version of posts that are attractive, colorful, and extensively being used among the social media community.

An internet meme is a viral concept that alters as it spreads out [2]. It is usually taken as a movie clip or scene with an added subtext that has a modified meaning. As memes contain both image and text information it becomes easier to understand the context. Memes are being used to spread news, sharing knowledge, emotion, ideas, talents, etc. It is also being used by government agencies and industry professionals to disseminate awareness programs, advertise their products, ideas, and so on. At the same time, memes are also extensively being used to spread wrong, hate, or offensive information targeting a particular community based on their color, appearance, gender, religious believes, and so on (Puranik et al., 2021; Hegde et al., 2021; Yasaswini et al., 2021; Ghanghor et al., 2021b,a). Such offensive memes that intend to insult individuals or groups are termed as *troll* memes. Trolling memes can create social media conflicts and influence certain events. It can cause loss of business or ruin somebody's life. Therefore, it is essential to identify such trolling memes and restrict their spread.

There is a number of works proposed in the literature for memes analysis. For example, (Du et al., 2020), proposed a study on the themes contained in the text of memes extracted from Twitter. (Du et al., 2020) performed experiments using a neural network model. They found that thirty percent of memes in their dataset had a political theme and were mostly shared by democrats than repub-

[1] https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media

[2] https://whatis.techtarget.com/definition/Internet-meme

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 243–248
April 20, 2021 ©2021 Association for Computational Linguistics

licans. (Zubaidah and Ardelia, 2018) presented a discourse analysis of memes where they found that to understand memes, knowing social content of it is necessary, along with the text and picture. They also concluded that memes can be used for multiple purposes such as humour, sarcasm, identifying feelings, and so on. (Sharma et al., 2020) presented a memotion analysis dataset for categorizing memes into annotated tags namely sentiment, and type of humor that is, sarcastic, humorous, or offensive. (Gomez et al., 2020) proposed a hybrid approach using inception network and long short term memory network for extracting features from image and text respectively to identify hate content in the multimodal dataset. (Suryawanshi et al., 2020a) proposed a multimodal meme dataset (MultiOFF) for offensive content detection. They used an early fusion technique to combine the image and text modality.

As discussed above, most of the existing works have proposed a hybrid approach for tackling image and text present in memes. Moreover, the text language used in those memes was English. The current paper is unique in the sense that it identifies offensive or trolling memes where texts are in the Tamil language. Tamil was the first to be categorized as India's classical language and is one of the world's longest-surviving classical languages. Over 55 percent of the epigraphic inscriptions discovered by the Archaeological Survey of India (about 55,000) are in the Tamil language. In Sri Lanka and on commercial products in Thailand and Egypt, Tamil language inscriptions written in Brahmi script have been found (Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2021a,b; Thavareesan and Mahesan, 2019, 2020a,b). The dataset used in the paper is a part of the task *EACL 2021:Meme classification for Tamil*. The problem proposed is a binary classification problem where every meme has to be categorized into tags *troll* and *not troll*. The classification model used for this was a hybrid model having 2-layered CNN and Bi-LSTM in parallel[3]. The proposed model on test data reported precision 0.52, recall 0.59, and weighted F1 0.3.

The rest of the article is organized as follows; Section 2 discusses the given task and dimension of the dataset. The methodology of the paper is explained in Section 3. This is followed by experimental results which are explained in Section 4. Finally, Section 5 concludes the paper.

2 Task and Data Description

As per the best of our knowledge, this is the first task of identifying offensive *troll* meme in Tamil. The task was proposed in *EACL 2021*. The objective of the task was to classify each meme into two categories *troll* and *not troll* (Suryawanshi and Chakravarthi, 2021; Suryawanshi et al., 2020b). Training and testing dataset of the task was released in a phased manner where initially, training data was released in a folder containing two files; a csv and an image folder. The image folder contained 2300 images out of that 1282 image names were starting with *troll* and other 1080 image names were starting with *not troll*. Based on their naming convention we identified which images belonged to *troll* class or *not troll* class. The csv file contained two fields, image names and their captions (or texts). Similarly, for the test dataset, two files (image folder and csv file) were given to predict the final class label. But, in the test set the image names were not having class information. The total test set count was 667.

3 Methodology

To identify *troll* meme in Tamil a hybrid binary classification model is proposed. The detailed flow of the proposed model is shown in Figure 1. The detailed working of the model is discussed in the subsections.

3.1 Data Preprocessing

The prepossessing steps were employed in both image and text parts in the given meme data. It was necessary to represent the images and texts into equal length vectors respectively to further use it as an input to the proposed model.

In the caption column of csv file, we checked for missing values and found that for a few images, the caption field was blank. Those missing values were replaced with "0". Then, we performed text cleaning by removing punctuation, extra space from the text. The cleaned texts were then tokenized and encoded into a sequence of token indices. Finally, to make all the texts of equal length, padding was performed with a maximum length of 25 (the average length of sentence in the dataset).

In the image folder, we fetched each image one by one and performed a few prepossessing steps.

[3] Developed model code repository can be found here: `https://github.com/ankitmishra2232/Troll-memes-classification-in-Tamil-`

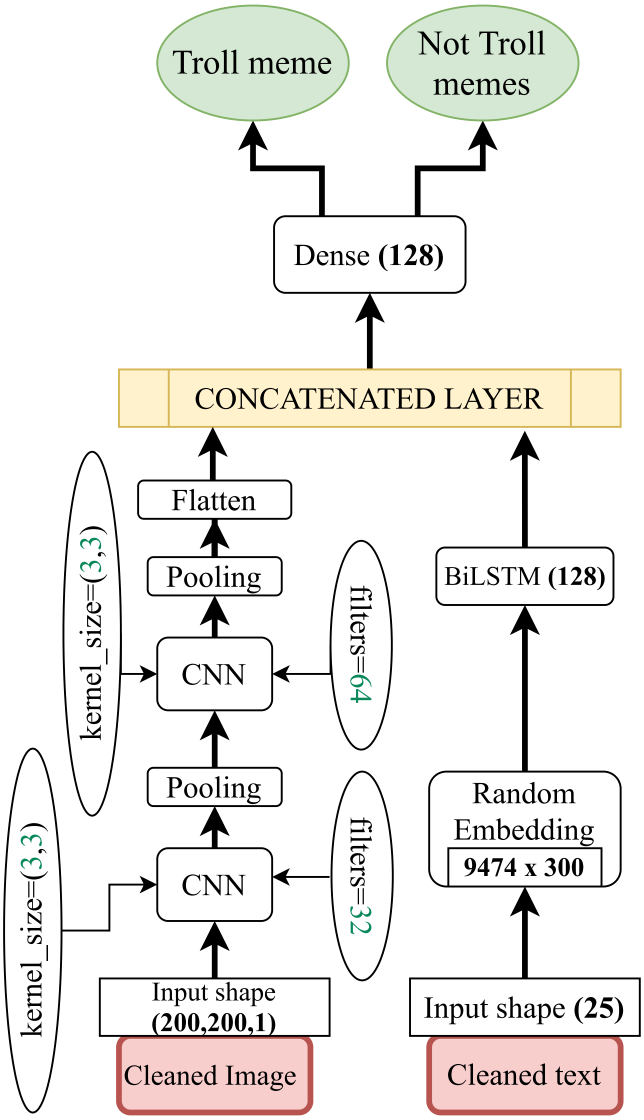
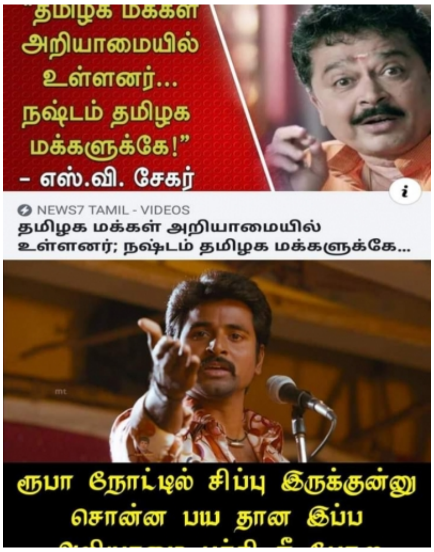


Figure 1: Proposed hybrid model for meme classification

First, we resized each image into an equal shape of a vector (200,200,3). Image processing usually requires a lot of resources. To avoid the resource scarcity problem we converted the colored images into gray-scale. Figure 2 and 3 shows an example of color image and gray-scale image. The gray-scale image was finally applied with denoising steps to make the image smoother and noise-free.



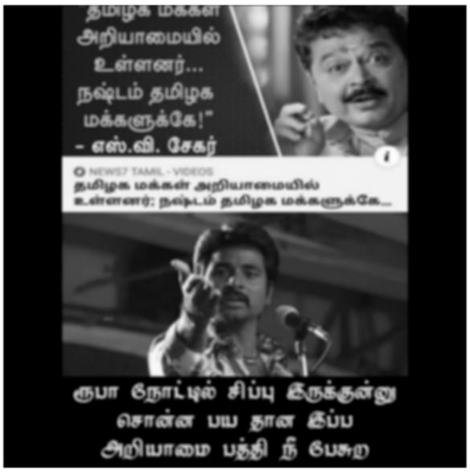


Figure 3: An example of gray-scale image from the dataset

3.2 Classification Model

A hybrid parallel deep learning model is developed to extract features from text and image using convolution neural network (CNN) and bidirectional long short term memory (Bi-LSTM) model respectively. The flow diagram of the model is shown in Figure 1. This section discusses the working of text and image models one after the other.

The input to the text model was image caption. For every word in the caption, a one-hot vector is created having a dimension equal to vocabulary size. In our dataset, the vocabulary size was 9474. Therefore, every word was represented in a 9474-dimensional vector. This high dimensional vector contained a single ‘1’ and rest all zeros. The position of ‘1’ in the vector was the word index number obtained while tokenizing the corpus. The input one-hot vector was then passed through an embedding layer that represented them in a 300-dimensional dense vector. The embedded vector obtained from each word was fed to the Bi-LSTM layer having 128 units. The output obtained at the last timestamp (for 25th word) of Bi-LSTM was considered as a feature vector of text.

The input to the image model was the matrix of the gray-scaled image obtained after preprocessing. The shape of the matrix was (200,200,1). The image matrix was passed to the two-layered CNN-pooling layer. In the first CNN layer, 32 filters were used having kernel size (3,3). They performed a two-dimensional convolution on the image matrix and resulted in a feature matrix that was passed through the ReLU activation function. The most prominent features were then pooled from the feature matrix in the window (2,2). The same process was repeated for the second CNN-pooling layer with 64 filters. The feature matrix obtained from the second CNN-pooling layer was converted into a flat vector and considered as the feature vector of the image.

The vectors obtained from image and text were then combined at concatenated layer. Finally, the combined feature was given input to a fully connected dense network which classified each input into *troll* and *not troll*. The training of the model was done for 30 epochs with adam optimizer and binary cross-entropy loss function.

4 Result

All models were developed in *Python* using libraries Sklearn, Keras, Pandas, Numpy, and so on. We experimented with different variations of the hybrid settings such as a parallel single-layer CNN-LSTM model, a parallel single layer CNN-CNN model, a parallel single-layer CNN-Bi-LSTM model, a parallel multilayer CNN-LSTM model, a parallel multilayer CNN-CNN model, and a parallel multi CNN layer and single Bi-LSTM layer model. We are reporting here the results of the best model, a parallel multi CNN layer and a single Bi-LSTM layer model for the test dataset. Initially, organizers provided test data without labels. The prediction of each test data was submitted to the competitions. Later, the organizers (*EACL 2021: Troll Meme classification in Tamil*) released the score of the testing data as 0.52 precision, 0.59 Recall, and 0.3 F1 scores. Our model obtained the 10th position in the ranking released by organizers. The 1st ranked model in the competition obtained a precision of 0.57, recall 0.6, and F1-score 0.55. As it is observed, the recall of the current model (0.59) was very close to the recall of the 1st ranked model (0.60). However, the precision of the submitted model (0.52) was a little far from the precision of the best-ranked model (0.59). That means the false positive rate of our model was higher.

5 Conclusion and Future work

The current paper identified *troll* Memes in social media platform using a parallel CNN-Bi-LSTM based network. The task was proposed by *EACL 2021* for *troll* Meme classification in Tamil. The proposed model categorised each meme into one of the classes *troll* and *not troll*. The model reported the precision of 0.52, recall 0.59, and F1-score 0.3 on the test dataset. Our model ranked at 10th position in the competition. The best model in the competition reported 0.57 precision, 0.6 recall, and 0.55 F1-score. That means a lot of works has to be done to improve the accuracy of the model. As future work, several transfer learning models can be employed such as VGG-16, ResNet50, inception network, and so on for image and BERT classifiers with a few pre-trained embedding models like GloVe, Word2Vec, FastText, and so on for text processing.

References

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldhana, John Philip McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021a. Findings of the shared task on Machine Translation in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021b. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164.

Nikhil Kumar Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. IIITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, Online. Association for Computational Linguistics.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja

Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Online.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. UVCE-IIITT@DravidianLangTech-EACL2021: Tamil Troll Meme Classification: You need to Pay more Attention. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-Hope Speech Detection: There is always hope in Transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Neneng Zubaidah and Irena Ardelia. 2018. A discourse analysis of memes. *Getsempena English Education Journal*, 5(2):58–64.